

构建 STKOS 术语发布与共享服务平台*

付鸿鹄[1] 张智雄[1] 刘建华[1,2] 钱 力[1,2] 王 颖[1]
[1](中国科学院文献情报中心 北京 100190)
[2](中国科学院大学 北京 100049)

摘要: 【目的】设计并实现 STKOS 术语发布与共享服务系统。【应用背景】作为一个超级词表, 科技知识组织体系(STKOS)需要提供给用户使用从而推进知识服务, 促进知识共享, 为此需通过一个共享服务平台对其进行发布。【方法】在对国际上相关项目和系统进行调研的基础上, 结合 STKOS 的特点和应用需求, 设计系统的功能框架, 并对系统实现中的关键问题包括应用场景、数据交换格式、数据结构、可视化、多版本管理等进行分析, 完成系统的整体建设。【结果】在千万量级数据场景下, 实现 STKOS 术语发布与共享服务系统平台。【结论】本系统支持 STKOS 数据的管理、发布, 支持对知识体系内容的揭示, 为用户提供对知识组织体系的浏览、检索和个性化定制下载。

关键词: 术语服务 科技知识组织体系 知识服务
分类号: G353.1

1 引 言

为了进一步加强我国科技信息资源与服务建设, 巩固和发展国家科技图书文献信息保障系统, 更有效地利用海量科技文献信息, 国家科技文献信息中心牵头组织实施了国家科技支撑计划“面向外文科技文献信息的知识组织体系建设和示范应用”项目[1]。项目在整合国外科技领域数百个相关词表、规范、本体的基础上, 构建科技知识组织体系(Science and Technology Knowledge Oranization System, STKOS), 形成一个超级词表。科技知识组织体系发布服务系统是该项目面向用户的服务系统, 提供科技术语发布与共享服务, 是一个面向最终用户的术语服务系统。

国外许多机构已经开展了术语服务系统的相关研究和建设, 如 OCLC 术语服务[2]、STAR (Semantic Technologies for Archaeological Resources) [3]、JISC-HILT[4]、UMLS 术语服务[5]、VocBench[6]等。已有论文对国外术语注册与术语服务进行详细介绍[7], 并对术语服务系统的功能和服务进行比较全面的总结, 此处不再详述。

本文主要目标是为共享和重用 STKOS 的研究与建设成果, 设计并实现一个科技知识组织体系的发布和服务平台。对国际上相关项目和系统进行深入调研, 在总结其经验和功能、技术特点的基础上, 结合 STKOS 的特点和应用需求, 完成系统平台的设计和开发。系统设计参考 UMLS 术语服务、VocBench 等国外比较成功的术语服务系统。系统构建在已经建成的 STKOS 基础上, 对外提供检索查询、浏览导航、详情展示等服务, 用户可以方便地获取、查阅和利用科技知识组织体系; 系统提供 STKOS 的下载功能, 可以按照标准格式输出学科领域内的 STKOS 片段, 以促进 STKOS 的共享和应用, 提供多种方式浏览和利用

通讯作者: 付鸿鹄, ORCID: 0000-0002-6642-2922, E-mail: fuhh@mail.las.ac.cn。
*本文系国家科技支撑计划课题“科技知识组织体系共享服务平台建设”子课题“科技知识组织体系(STKOS)发布服务系统”(项目编号: 2011BAH10B03-2)的研究成果之一。

STKOS 及 STKOS 子集，以满足不同机构、不同用户对领域知识本体的需求。STKOS 术语发布与共享服务系统平台的建设，实现对跨学科的、千万数量级的超级科技词表的共享和发布，实现对词表不同视角的内容揭示(多视图的结果揭示、多维可视化分析)，为用户提供定制和下载，促进词表内容共享，达到最大程度地共享和重用科技知识组织体系成果的目的。

2　系统功能框架

STKOS 术语发布与共享服务平台的总体建设目标是在已经建成的科技知识组织体系 STKOS 和基于 STKOS 的知识本体引擎的基础上，构建 STKOS 的术语发布与共享服务。对外提供检索查询、浏览导航、详情展示服务，让用户可以方便地获取、查阅和利用科技知识组织体系。同时，为了促进 STKOS 更为广泛的应用，提供 STKOS 的定制与集成，包括输入、输出、下载和提供第三方知识组织体系的集成功能。STKOS 术语发布与共享服务平台的后台管理系统主要是对 STKOS 的各种版本进行管理，对用户的各种操作权限进行管理，提供对 STKOS 的增加、编辑等维护工具。

根据系统的建设目标，STKOS 术语发布与共享服务平台的整体框架包括前台服务系统、定制与集成系统和后台管理系统三个模块。系统总体框架如图 1 所示:

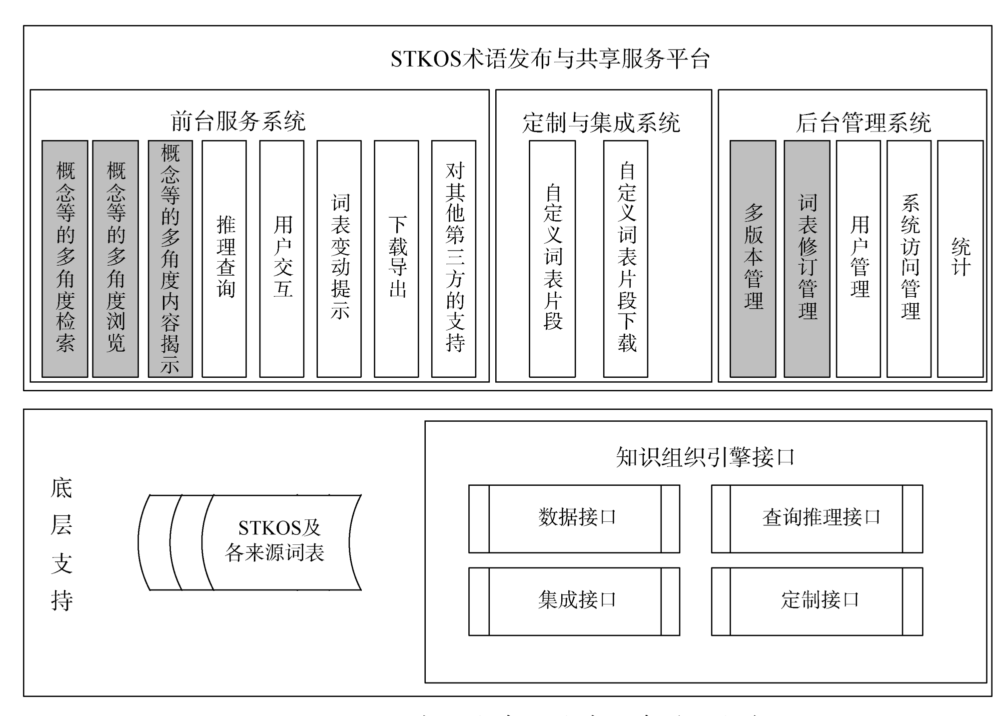


图 1　STKOS 术语发布与共享服务系统框架

(1) 前台服务系统主要与用户交互，向用户提供各类显性化的知识服务，对外提供检索查询、浏览导航、详情展示服务，用户可以方便地获取、查阅和利用科技知识组织体系。通过前台服务系统，用户可以获取所需的数据信息。

(2) 定制与集成系统支持用户自定义知识组织体系的功能，用户可利用此模块自主选择定制所需的知识组织体系实现个性化的定制服务，自定义用户所需的知识组织片段(知识可来源于多个词表、多个范畴)；自定义知识组织片段的多格式下载。可能的应用场景包括：用户需要排除不需要的词表或在本地应用程序不可使用的词表、用户需要使用多种数据输出选项和过滤器自定义的一个子集等。

(3) 后台管理系统主要负责用户权限管理、用户操作行为管理、系统中各版本的知识组织体系发布与组织管理以及知识组织体系中修订内容的管理等。

STKOS 术语发布与共享服务系统构建在底层 STKOS 与各来源词表、知识组织引擎基础上。整个服务平台对外服务时将采用账户登录的方式进行权限的控制，依据用户角色的不同，区分不同的服务。系统数

据流如图 2 所示:

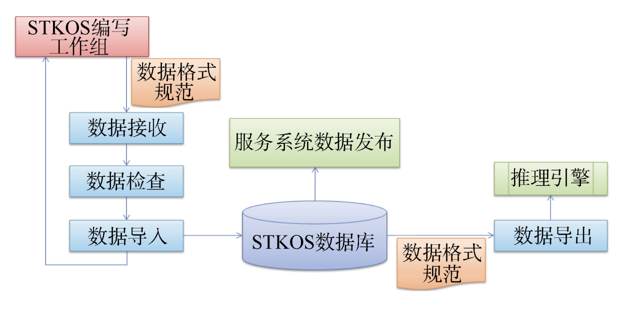


图 2　系统数据流

3　系统设计与实现中的关键问题

3.1　应用场景分析

STKOS 术语发布与共享服务平台提供基于科技超级词表的术语服务。除机器用户访问接口外, 重点考虑基于 Web 浏览的人类用户的查询访问需求。从人类用户的应用场景来看, 将用户区分为普通用户和情报学领域专家两种类型。根据两类用户的知识背景和使用目标的差异, 提供差异化的服务界面。

3.2　数据交换格式

术语发布与共享服务平台在 STKOS 的基础上提供术语发布服务, 本身并不加工生产 STKOS 数据, 需从 STKOS 编写工作组接收具体的数据内容。接收数据的格式和质量直接影响 STKOS 的服务。STKOS 发布元数据定义了接收数据规范和从 STKOS 编写工作组接收的 STKOS 数据内容的具体格式(Schema)。数据接收模块根据数据接收 Schema 定义的元数据, 对接收的数据进行解析并导入术语发布与共享服务平台。STKOS 发布元数据定义了 6 大类基本数据模型[8], 包括来源术语、来源词表、科技术语、STKOS 规范概念、范畴类和范畴表, 并对基本数据类型之间的关系进行详细描述, 如图 3 所示:

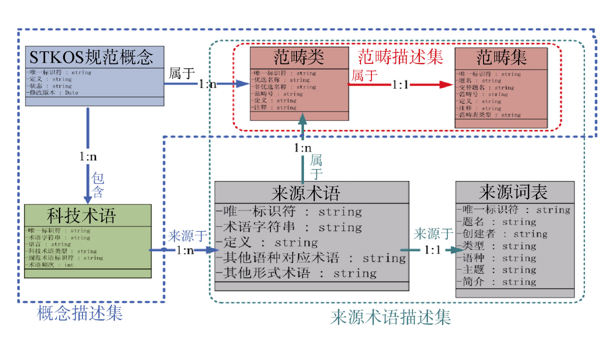


图 3　发布元数据 Schema 基本结构[8]

3.3　基本信息数据结构

STKOS 数据整合了科技领域数百个来源词表, 包含来源词表中的数据信息及整合、加工的数据信息, 具有数据来源多样、数据量大、具有上层范畴、存在多种类型的数据关系的特点。

数据结构的设计需考虑 STKOS 数据自身的特点、术语发布服务上的需求, 并需遵循相关标准规范。因此在设计上主要参考 ISO 25964-1 数据模型、STKOS 发布元数据、UMLS 数据模型。ISO 25964-1 数据模型定义了叙词表相关标准规范, 提供数据模型建议; STKOS 发布元数据提供 STKOS 的数据模型, 其设计主要参考 ISO 25964-1 和 UMLS 数据模型; UMLS 整合医学领域的上百个词表, 其数据在内容和形式上与 STKOS 有许多相似之处, 具有借鉴意义。

数据模型的设计思路是参考 ISO 25964-1 数据模型, 遵循 STKOS 发布元数据规范, 借鉴 UMLS 的相关设计思想, 允许数据的适当冗余。STKOS 发布元数据规范中定义了来源术语、基础术语、概念三个层次的数据类型, 术语和概念的关系主要建立在来源术语的基础之上, 体现来源术语之间的关系, 术语概念关系主要建立在概念和来源术语的基础之上。

STKOS 发布元数据定义了 6 大类基本数据模型, 包括来源术语、来源词表、科技术语、STKOS 规范概念、范畴类和范畴表。在此基础上, 结合发布服务系统的实际应用需求, 重点设计规范概念–科技术语–来源术语三层数据结构, 以及各种概念、术语关系数据模型。

3.4　数据可视化

STKOS 中存在多种类型的语义关系, 如术语–术语关系、术语–概念关系、术语–范畴关系、概念–概念关系、概念–范畴关系以及范畴–范畴关系等类型。而在 STKOS 中, 概念与术语是知识主体, 概念间关系又主要通过术语间关系体现。知识组织体系中的可视化揭示是展示知识组织内部相关关系的重要手段。STKOS 由多个词表体系组成, 有高层范畴规范, 具有词表间概念映射、同义词、优选词、各种关系类型等特点, 因此如何有效揭示一个概念和其他相关概念的各种关系是一个复杂问题。

在知识组织体系的可视化基础上, 结合 STKOS 特点, 构建一个 STKOS 多维语义可视化揭示模型, 从

术语本身关系、相关关系、层级关系、语义关系、演化关系等多个维度对数据关系进行深入揭示。这种基于多维度的 STKOS 概念关系可视化揭示方法可以有效揭示 STKOS 中一个概念与其他概念的相关关系，进而揭示整个知识组织体系内在丰富的逻辑关联关系，使用户通过各种可视化关系深入了解某一概念在整个 STKOS 中所处的位置，帮助用户理解和使用整个知识组织体系。选择 Cytoscape Web[9]与 D3.js[10]作为可视化的基础工具并进行二次开发。

3.5　数据的版本管理

多版本数据的发布是 STKOS 中数据版本演变分析和版本更新提示的基础。版本管理对各种不同类型版本的 STKOS 进行导入、浏览、存档、导出等操作，为用户跨版本应用 STKOS 数据提供支持。通过 STKOS 版本管理，可以回溯不同时期、不同版本 STKOS 的变迁；也可及时反映当前 STKOS 版本内的修改、变动情况，为 STKOS 发布服务系统各个功能模块提供跨版本的数据接口。

设计主要参考了 UMLS 版本历史变更记录文件格式、联合国粮食及农业组织(FAO)农业词表管理平台 VocBench 中词表最新变更表。UMLS 版本历史变更记录文件反映了 UMLS 正式版本针对前一版本的重大变更记录。而 VocBench 中则通过词表最新变更表记录同一版本中词表修订情况。两者结合起来，恰好可以满足 STKOS 后台管理系统既要保存正式发布的多版本数据，又要及时反映当前最新版本变更信息的需求。设计要求为每一个 STKOS 历史版本信息完整记录存档，作为长期保存原始记录，且不能修订。当前服务版本是在最近一期历史版本的基础上，及时增加加工平台当期最新的修订信息。变更记录表应能完整反映历史版本之间以及当前服务活动版本的各种信息变更情况。

3.6　构建多核索引

STKOS 发布服务具有数据量大、关系复杂、检索多维度的特点。为了满足系统快速响应的需求，需要对索引体系进行优化处理。系统索引基于 Solr 构建。基于对数据查询关系的分析，设计基础数据索引(包括概念索引、术语索引)、关系索引(包括相关关系索引、层级关系索引、共现关系索引)共两类 5 种索引以支持对概念的多角度检索查询。基础数据索引包括范畴、概念、术语、原子、属性等基本信息。

4　系统实现效果分析

4.1　检索及结果展示

检索是前台服务系统的核心模块。检索目标对象主要为概念术语，用户可根据需要选择特定的检索方式，设定检索条件查询所需信息。提供关联检索，即检索与某指定术语词/概念存在特定关系的概念术语。提供增强或扩展检索，通过术语关系、关联信息提供与检索词存在语义关联的其他概念提示。

用户可以从多个方面限制检索条件，系统提供检索提示功能。检索结果中对检索词进行高亮显示，并对近期更新的条目显示更新提示。针对检索结果集提供按列表展示、按范畴分面、按来源词表分面三种展示形式。针对检索结果详细信息提供文本视图、RDF 视图、图形化视图等多种视图。系统检索结果详情展示界面如图 4：



图 4　检索结果详情

4.2　检索结果多维可视化

检索结果的多维可视化使用户可以从多个维度全面了解检索目标及其在整个知识体系中的位置。可视化结果从词本身关系(即术语自身重要属性关系)、相关关系、层级关系、语义关系、演化归并关系等维度实现。图 5 为多维可视化结果呈现界面。用户可以通过选择不同的标签在不同维度的视图之间进行切换。

4.3　浏览与个性化定制

提供按范畴体系的浏览。范畴体现整个知识组织体系的学科领域结构，通过范畴浏览展示，用户可以

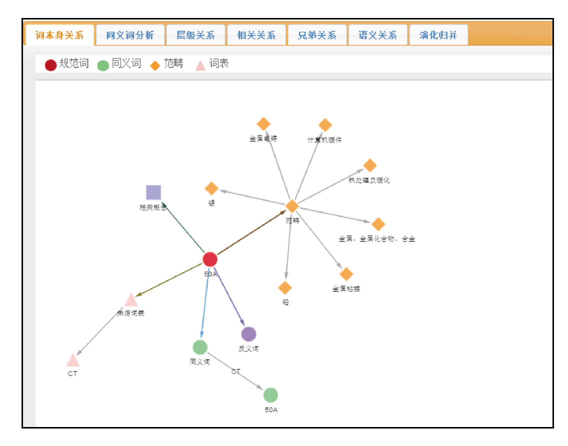


图 5　多维可视化结果

快速定位到自己感兴趣的领域，了解领域的整体知识结构和知识内容的概括信息。此外，为了方便用户使用，系统提供基于范畴浏览实现对子集的定制，定制后生成所选范围内概念、术语及相关信息的子集，并提供用户下载。图 6 为按范畴浏览并进行子集定制的用户界面。用户可以通过范畴导航和首字母导航选择感兴趣的内容进行子集定制。

图 6　范畴浏览与定制

5　结　语

STKOS 术语发布与共享服务平台是实现 STKOS 建设成果共享的重要平台，对于知识组织和知识服务具有重要意义。结合系统建设过程中的具体问题，在构建术语服务时，需要注意以下几个方面:

(1) 数据交换格式。数据交换格式包括两个方面，即术语服务系统接收数据的格式和进行数据发布的格式。接收数据格式的设计必须基于词表加工系统数据格式，兼顾发布服务系统对数据内容揭示的需求及效率上的需求，做到规范性和灵活性。发布数据格式需考虑用户与第三方系统的需求，考虑数据内容与其他知识组织体系之间的映射关系，应遵循数据交换标准和规范。

(2) 底层数据模型。由于 STKOS 整合了科技领域上百个词表，数据来源多样，并在上层构建了范畴体系，进行了术语概念的归并处理等。底层数据模型必须能够完整保留各类信息内容，从而保证对各类知识内容的完整揭示。

(3) 随着知识领域的发展，相关的术语、概念等也随之变化，构建在知识组织体系基础上的术语服务系统必须充分考虑对未来的兼容性。知识组织体系内容的多版本管理为数据版本演变分析和版本更新提示提供数据基础。

(4) 可视化是帮助用户理解、认识知识组织体系中各种数据对象及其关系的一种直观形式，是展示知识组织内部相关关系的重要手段。可视化形式要符合人们的认知习惯，易于理解，从而有效揭示一个概念和其他相关概念的各种关系。

此外，目前的术语服务以规范后的 STKOS 超级词表为主要内容，STKOS 超级词表是一个受控词表，但在实际服务中，非受控词表对用户的意义重大，如何在受控词表与非受控词表之间建立关联，从而实现相应的服务也是系统未来建设中需要考虑的一个重要问题。

参考文献:

[1] 科技知识组织体系共享服务平台建设[EB/OL]. [2015-03-03]. http://sharestkos.las.ac.cn/. (Construction of STKOS Sharing Service Platform [EB/OL]. [2015-03-03]. http://sharestkos.las.ac.cn/.)

[2] Terminology Services [EB/OL]. [2015-03-03]. http://www.oclc.org/research/projects/termservices/.

[3] Semantic Technologies for Archaeological Resources [EB/OL]. [2015-03-03]. http://hypermedia.research.southwales.ac.uk/kos/star/.

[4] JISC-HILT [EB/OL]. [2015-03-03]. http://pure.strath.ac.uk/portal/files/384452/strathprints014046.pdf.

[5] Unified Medical Language System [EB/OL]. [2015-03-03]. http://www.nlm.nih.gov/pubs/factsheets/umls.html.

[6] VocBench [EB/OL]. [2015-03-03]. http://aims.fao.org/vest-registry/tools/vocbench-2.

[7] 欧石燕. 国外术语注册与术语服务综述[J]. 中国图书馆学报, 2014, 40(5): 110-126. (Ou Shiyan. A Review of Foreign Terminology Registries and Terminology Services [J]. Journal

of Library Science in China, 2014, 40(5): 110-126.)

[8] 宋文. STKOS发布元数据设计[EB/OL]. [2015-03-03]. http://168.160.16.186/conference/dome_ch/2012/downloads/pdf发言稿/宋文_STKOS发布元数据设计.pdf. (Song Wen. Design of STKOS Publishing Metadata [EB/OL]. [2015-03-03]. http://168.160.16.186/conference/dome_ch/2012/downloads/pdf发言稿/宋文_STKOS发布元数据设计.pdf.)

[9] Cytoscape Web [EB/OL]. [2015-03-03]. http://www.cytoscape.org/index.html.

[10] D3.js [OL]. [2015-03-03]. http://d3js.org/.

作者贡献声明：

付鸿鹄：系统详细设计，数据分析，部分系统开发，论文撰写；
张智雄：系统框架设计；
刘建华：前期调研，系统框架设计；
钱力：系统详细设计，系统开发；
王颖：数据分析，数据交换格式设计。

收稿日期: 2015-03-03
收修改稿日期: 2015-03-31

Construction of STKOS Term Publishing and Sharing Service Platform

Fu Honghu[1] Zhang Zhixiong[1] Liu Jianhua[1,2] Qian Li[1,2] Wang Ying[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] Design and implement the STKOS term publishing and sharing service platform. [**Context**] As a metathesaurus, STKOS needs be published to public/organization users and promotes knowledge service and sharing. [**Methods**] Based on the research of international projects and systems about term service, analysing the features and requirements of STKOS, the framework of the system is designed and implemented. The key issues are discussed, including application scenarios, data exchange formats, data structures, visualization, multi-version management, etc. [**Results**] Under the scenario of magnitude data, STKOS term publishing and sharing service platform is developed. [**Conclusions**] The system can support STKOS data management and release, contents revealing in STKOS, and browsing, retrieval and customized download for users.

Keywords: Term service STKOS Knowledge service